STA291

THURSDAY, 17 SEPTEMBER 2009

Administrative

- Suggested problems from the textbook

(not graded): 6.4, 6.5, and 6.6

We are between Ch 4 and Ch 6.

- Check MyStatLab for online homework

- Start bringing calculators (including labs—good to check skills!)

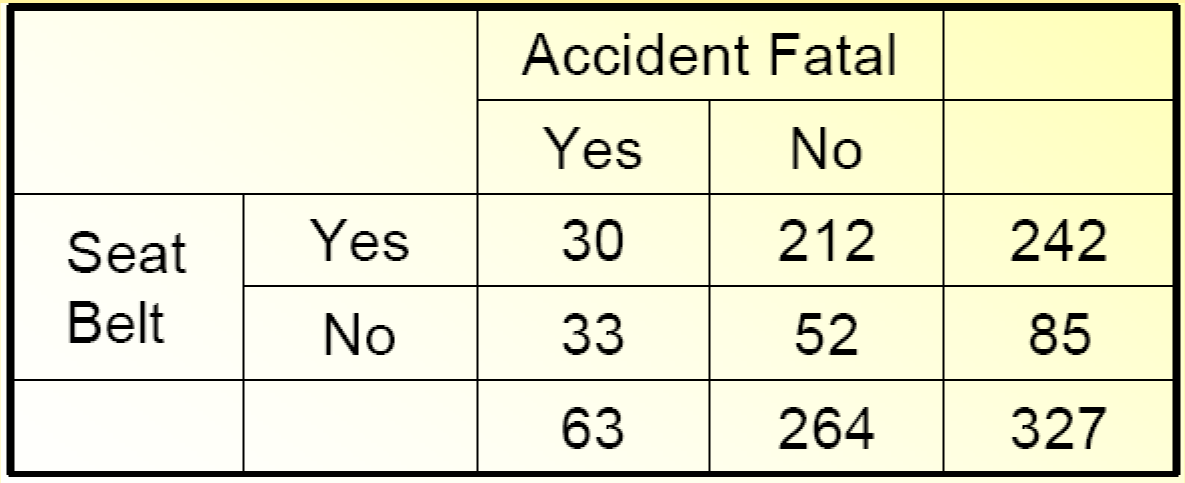
Where we've been …

- Data types (scales of measurement, etc.)
- Sampling methods (~~good, bad, ugly~~ SRS, stratified, cluster versus convenience, volunteer)—why is one group good and the other bad?

- Order we've covered these topics are the same order we would deal with these issues in a real-world problem

2 x 2 Contingency Table: Example

- 327 commercial motor vehicle drivers who had accidents in Kentucky from 1998 to 2002
- Two variables:
  - wearing a seat belt (y/n)
  - accident fatal (y/n)



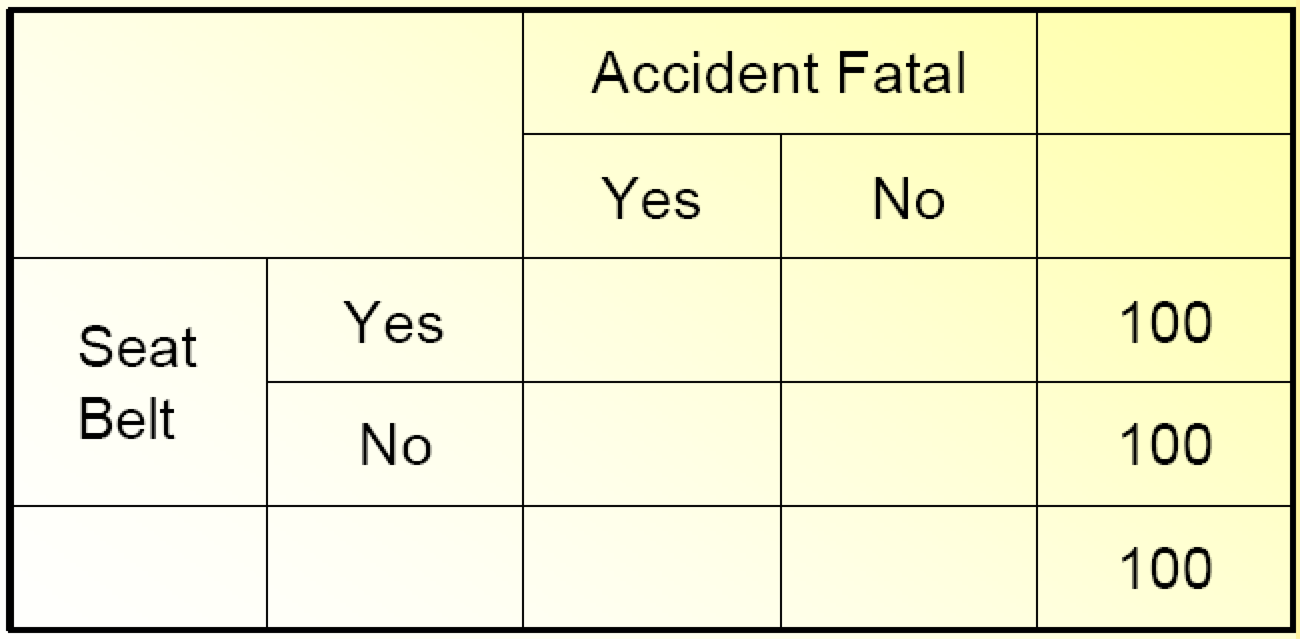
		Accident Fatal		
		Yes	No	
Seat Belt	Yes	30	212	242
	No	33	52	85
		63	264	327

2 x 2 Contingency Table: Example, cont’d.

- How can we compare fatality rates for the two groups?
- Relative frequencies or percentages within each row
- Two sets of relative frequencies (for *seatbelt=yes* and for *seatbelt=no)*, called **row relative frequencies**
- If seat belt use and fatality of accident are related, then there will be differences in the row relative frequencies

Row relative frequencies

- Two variables:
  - – wearing a seat belt (y/n)
  - – accident fatal (y/n)



		Accident Fatal		
		Yes	No	
Seat Belt	Yes			100
	No			100
				100

Describing the Relationship Between Two Interval Variables

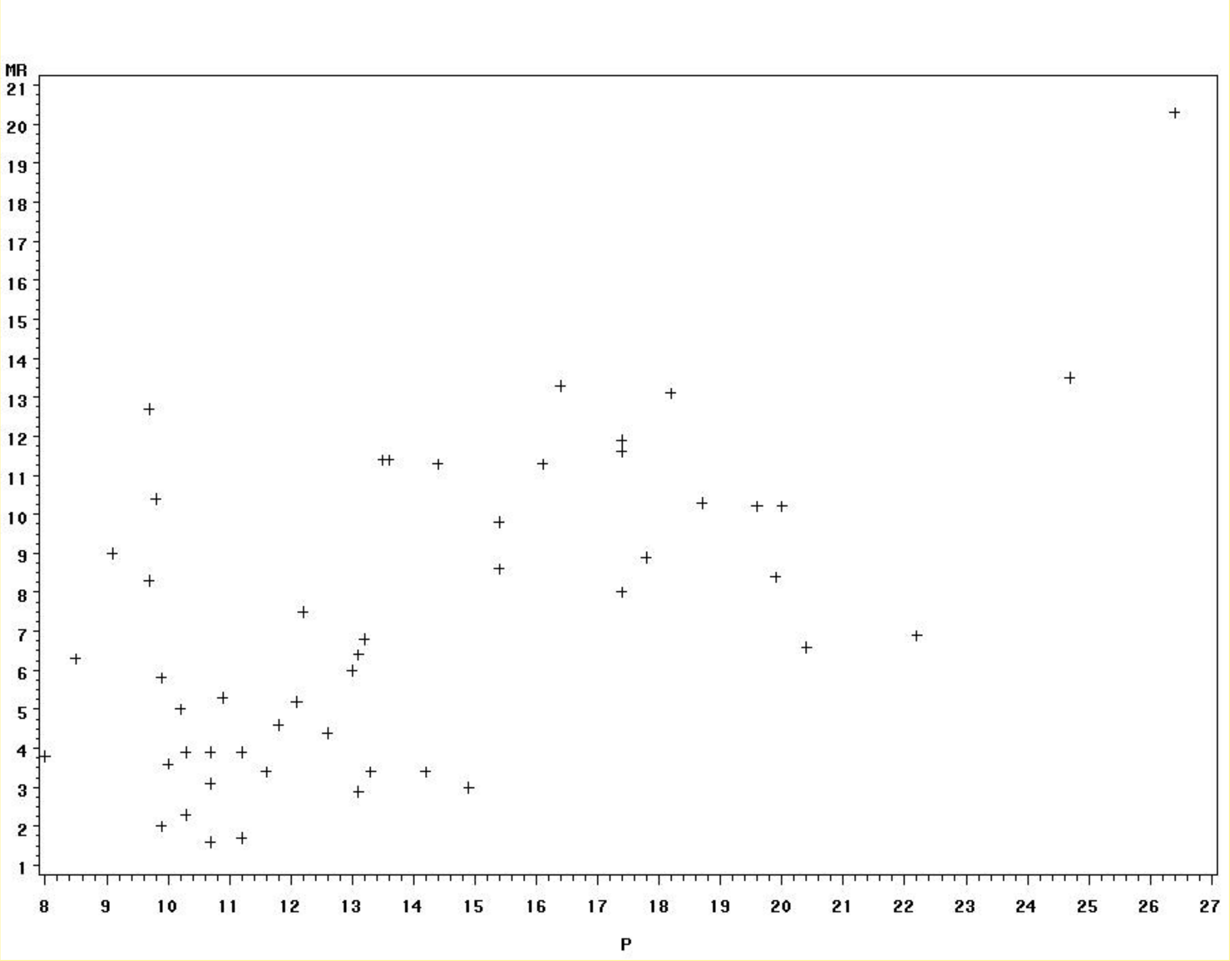
Scatter Diagram

- In applications where one variable depends to some degree on the other variables, we label the dependent variable *Y* and the independent variable *X*
- Example:

  Years of education = *X*

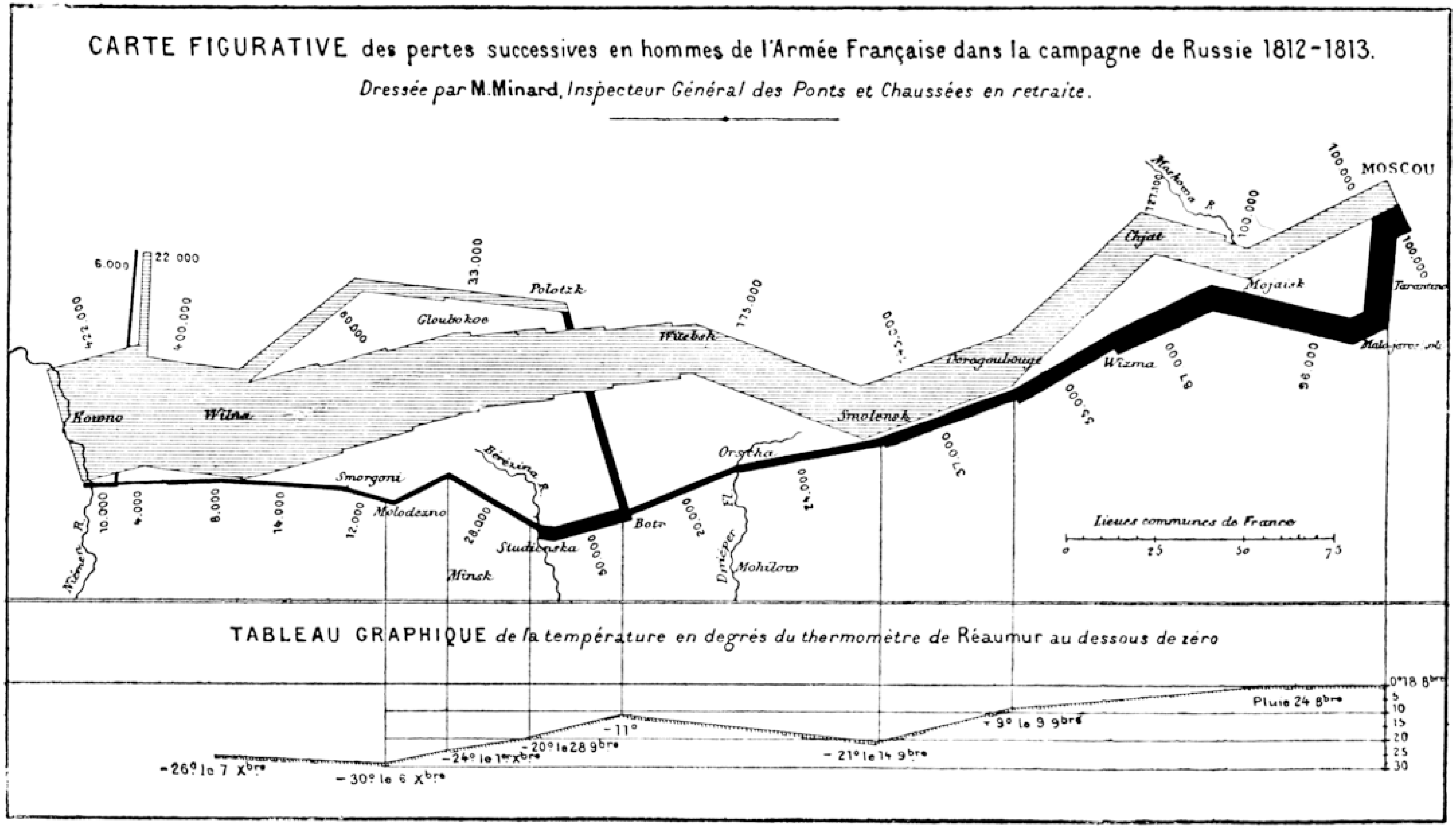
  Income = *Y*
- Each point in the scatter diagram corresponds to one observation

Scatter Diagram of Murder Rate (Y) and Poverty Rate (X) for the 50 States



3.1 Good Graphics …

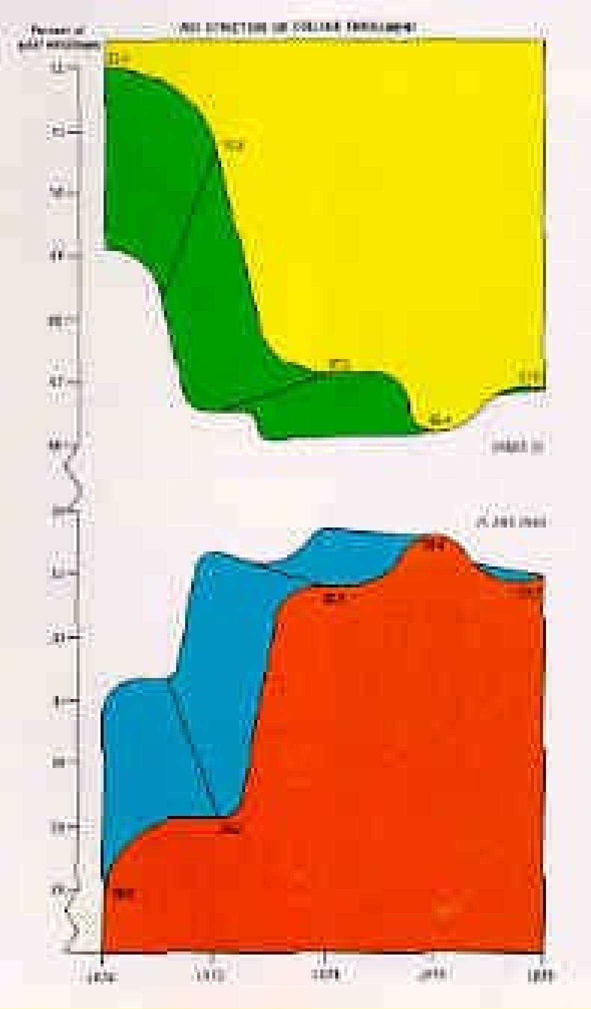
- … present large data sets concisely and coherently
- … can replace a thousand words and still be clearly understood and comprehended
- … encourage the viewer to compare two or more variables
- … do not replace substance by form
- … do not distort what the data reveal
- … have a high “data-to-ink” ratio


CARTE FIGURATIVE des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
MOSCOU
100.000
100,000
100.000
127.100
Moskowa R.
Chjat
Mojaisk
Tarantino
Malo-jarosewli
96.000
87.000
Wizma
55.000
Dorogobouge
37.000
Smolensk
145.000
175.000
Witebsk
33.000
Polotzk
Gloubokoe
60.000
422.000
400.000
6.000
22.000
Kowno
Wilna
Smorgoni
Molodezno
Bérézina R.
Studienska
Minsk
Botr
Orscha
Dnieper Fl.
Mohilow
Niémen R.
10.000
4.000
8.000
14.000
12.000
28.000
50.000
20.000
24.000
Lieues communes de France
0
25
50
75
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro
0° 18 8bre
Pluie 24 8bre
+9° le 9 9bre
-21° le 14 9bre
-11°
-20° le 28 9bre
-24° le 1er Xbre
-30° le 6 Xbre
-26° le 7 Xbre
5
10
15
20
25
30

3.2 Bad Graphics…

- …don't have a scale on the axis
- …have a misleading caption
- …distort by stretching/shrinking the vertical or horizontal axis
- …use histograms or bar charts with bars of unequal width
- …are more confusing than helpful

Bad Graphic, Example



Where next?

- **6 Numerical Descriptive Techniques**

– Review:

- **Parameter**

– numerical characteristic of the **population**

– calculated using the whole population

- **Statistic**

– numerical characteristic of the **sample**

– calculated using the sample

Measures of Central Location

- Also called Central *Tendency*
- "What is a typical measurement in the sample/population?"
  - Mean: Arithmetic average
  - Median: Midpoint of the observations when they are arranged in increasing order
  - Mode: Most frequent value

Mean (Average)

- Mean (or Average): Sum of measurements divided by the number of subjects

- Example: Observations 3,8,19,12

Mean =

Mathematical Notation: Sample Mean

- Sample size *n*
- Observations x_1 , x_2 ,..., x_n
- Sample Mean "*x*-bar"

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\Sigma = \text{SUM}$$

Mathematical Notation: Population Mean

- Population size N
- Observations x_1 , x_2 ,…, x_N
- Population mean μ (*mu*, read “myew”)

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Mean (Average)

- The mean requires numerical values. Only appropriate for quantitative data.
- It does not make sense to compute the mean for nominal variables.

- Example "Nationality" (nominal):

  Germany = 1, Italy = 2, U.S. = 3, Norway = 4

  Sample: Germany, Italy, Italy, U.S., and Norway
- Mean nationality = 2.4???

Mean (continued)

- Sometimes, the mean is calculated for ordinal variables, but this does not always make sense.
- Example "Weather" (on an ordinal scale):

  Sun=1, Partly Cloudy=2, Cloudy=3,

  Rain=4, Thunderstorm=5

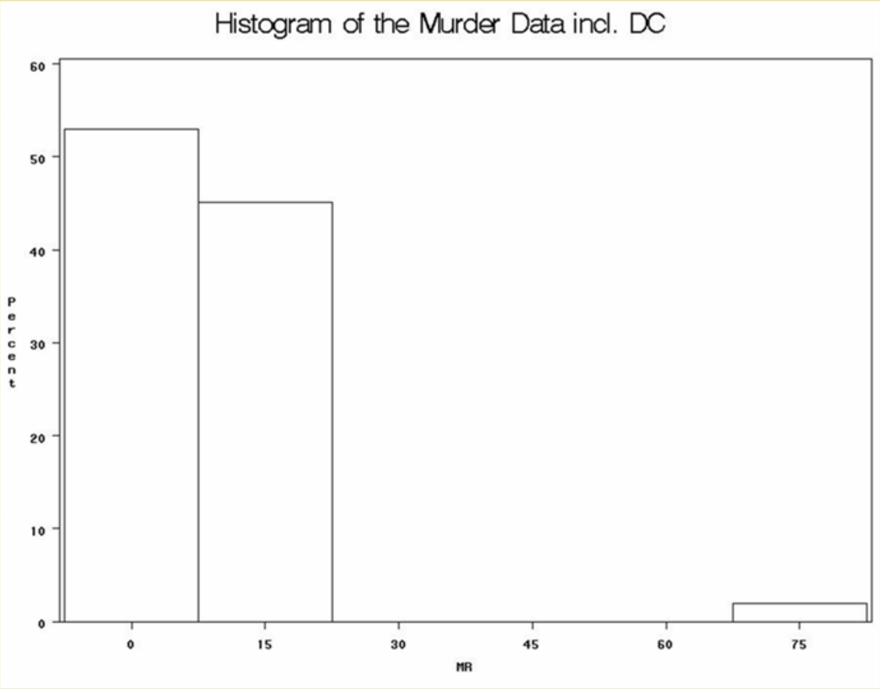
  - Mean (average) weather=2.8

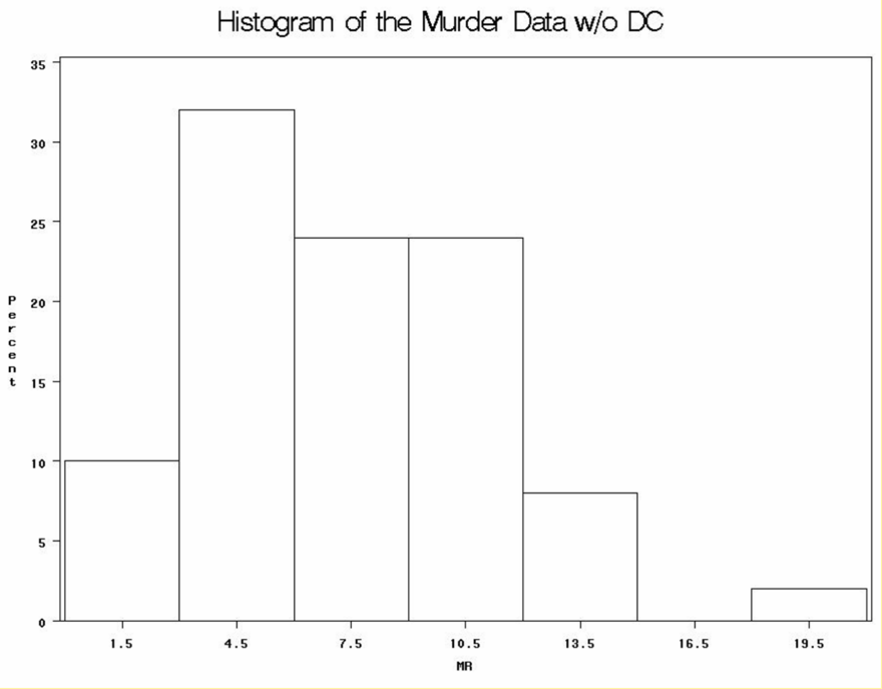
- Another example: "GPA = 3.8" is also a mean of observations measured on an ordinal scale

Mean(continued)

- The mean is highly influenced by outliers. That is, data points that are far from the rest of the data.

Example: Murder rates



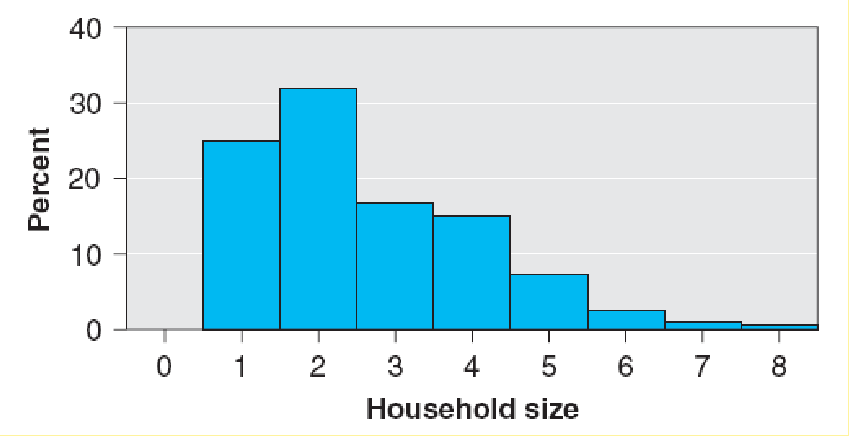


Mean (continued)

- Example: Murder Rate Data

Mean incl. DC: 8.73

Mean w/o DC: 7.33



- Any right-skewed distribution: the mean is “pulled” to the right

Central Location

- If the distribution is highly skewed, then the mean is not representative of a typical observation
- Example:

  Monthly income for five persons

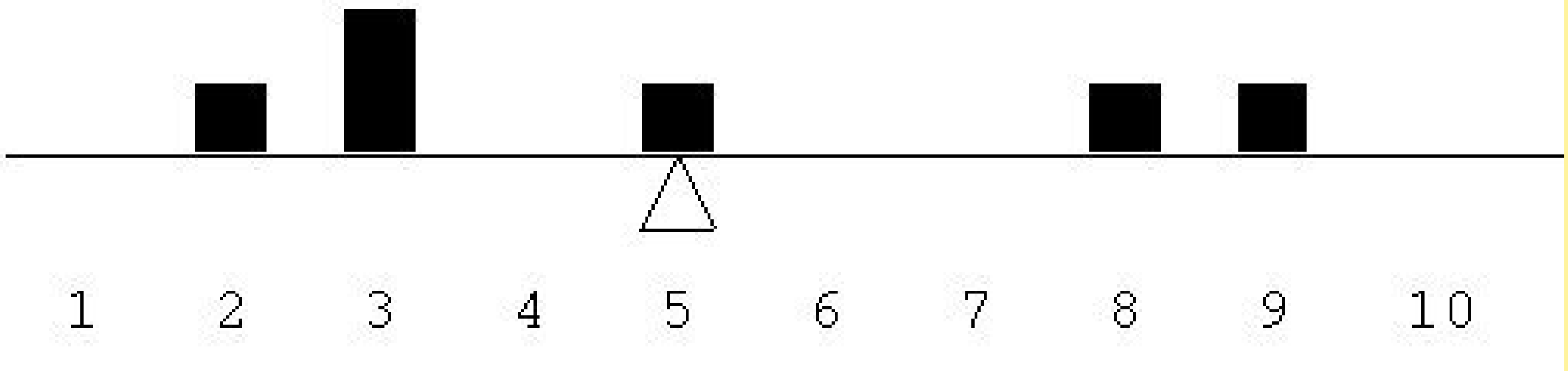
  1,000 2,000 3,000 4,000 100,000

  Average monthly income:

- Not representative of a typical observation.

Physical Interpretation of the Mean

- Assume that each measurement has the same “weight”



- Then, the mean is the center of gravity for the set of observations
- This is because the sum of the distances to the mean is the same for the observations above the mean as for the observations below the mean

Median

- The median is the measurement that falls in the middle of the ordered sample
- When the sample size *n is odd, there is a* middle value
- It has the ordered index *(n+1)/2*
- Example: 1.1, 2.3, 4.6, 7.9, 8.1

  n=5, *(n+1)/2=6/2=3,* *Index =3*

So,

Median = 3rd smallest observation = 4.6

Median

- When the sample size, *n*, is even, *average* the two middle values

- Example: 3, 4, 7, 10, 13, 19

n=6, $(n+1)/2=7/2=3.5$, Index =3.5

Median = midpoint between 3rd and 4th smallest observations = (7+10)/2 =8.5

Mean and Median

- For skewed distributions, the median is often a more appropriate measure of central tendency than the mean
- The median usually better describes a “typical value” when the sample distribution is highly skewed

- Example:

Monthly income for five persons (n = 5)

1,000 2,000 3,000 4,000 100,000

- Median monthly income: 3000

Mean and Median

- Example: Murder Rate Data

- Mean including DC: 8.73

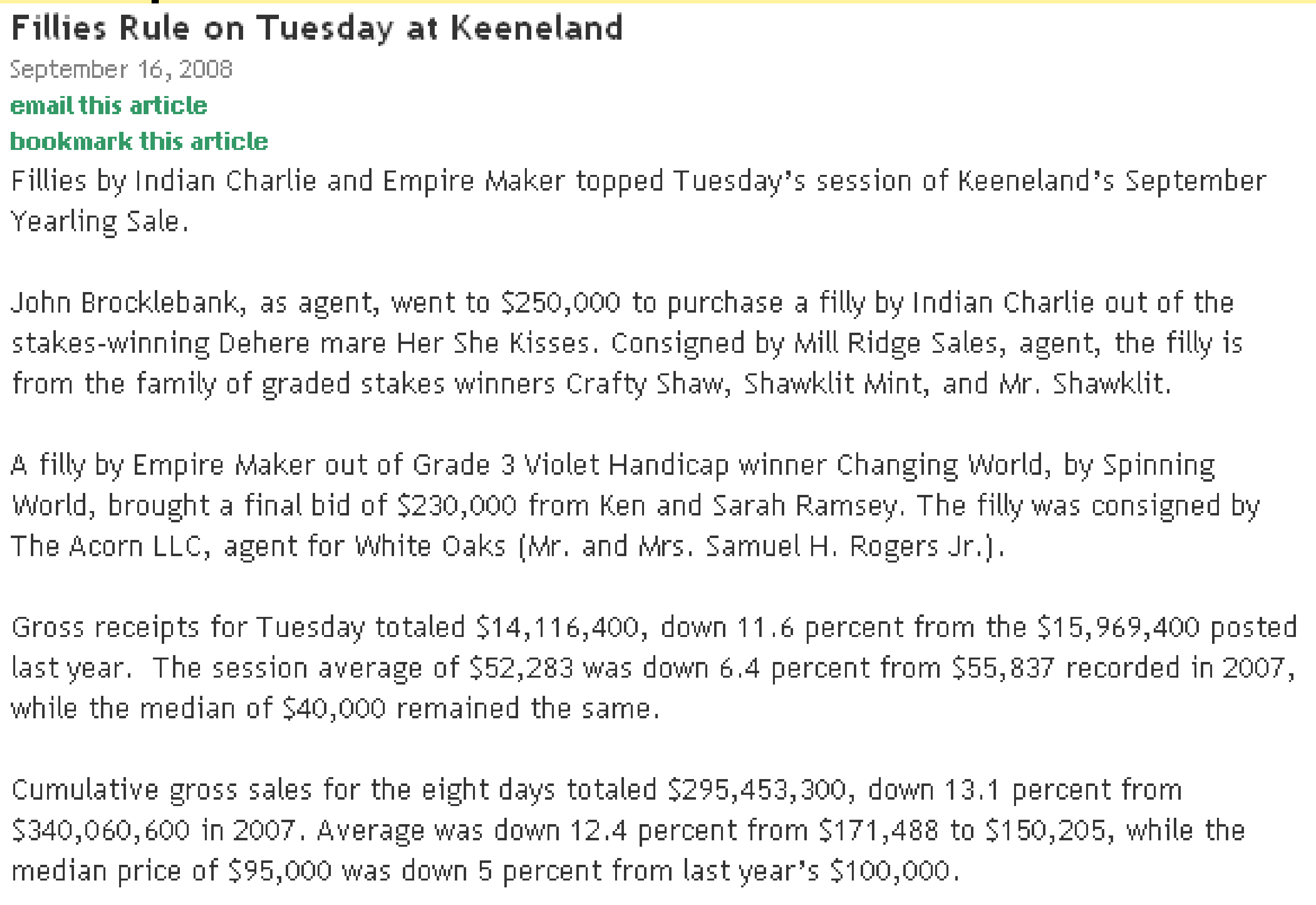
  Mean without DC: 7.33

- Median including DC: 6.8

  Median without DC: 6.7

Mean and Median

- Example: Keeneland Sales



Fillies Rule on Tuesday at Keeneland

September 16, 2008

email this article

bookmark this article

Fillies by Indian Charlie and Empire Maker topped Tuesday's session of Keeneland's September Yearling Sale.

John Brocklebank, as agent, went to $250,000 to purchase a filly by Indian Charlie out of the stakes-winning Dehere mare Her She Kisses. Consigned by Mill Ridge Sales, agent, the filly is from the family of graded stakes winners Crafty Shaw, Shawklit Mint, and Mr. Shawklit.

A filly by Empire Maker out of Grade 3 Violet Handicap winner Changing World, by Spinning World, brought a final bid of $230,000 from Ken and Sarah Ramsey. The filly was consigned by The Acorn LLC, agent for White Oaks (Mr. and Mrs. Samuel H. Rogers Jr.).

Gross receipts for Tuesday totaled $14,116,400, down 11.6 percent from the $15,969,400 posted last year. The session average of $52,283 was down 6.4 percent from $55,837 recorded in 2007, while the median of $40,000 remained the same.

Cumulative gross sales for the eight days totaled $295,453,300, down 13.1 percent from $340,060,600 in 2007. Average was down 12.4 percent from $171,488 to $150,205, while the median price of $95,000 was down 5 percent from last year's $100,000.

Mean and Median

- Is there a compromise between the median and the mean? Yes!
- Trimmed mean:

  1. Order the data from smallest to largest

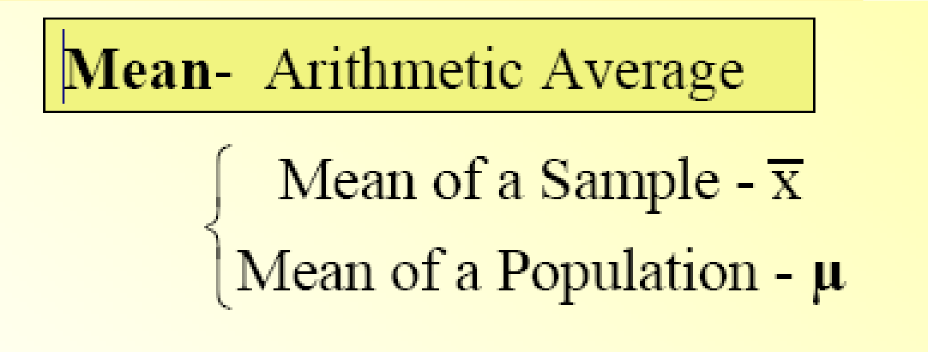
  2. Delete a selected number of values from each end of the ordered list

  3. Find the mean of the remaining values
- The trimming percentage is the percentage of values that have been deleted from each end of the ordered list.

Mode

- Mode of a data set is the most frequently occurring value
- Can speak of a data set being *unimodal*, *bimodal*, or *multimodal*
- Can be calculated on nominal (!) data
- On a histogram, where would the mode be?

Summary: Measures of Location

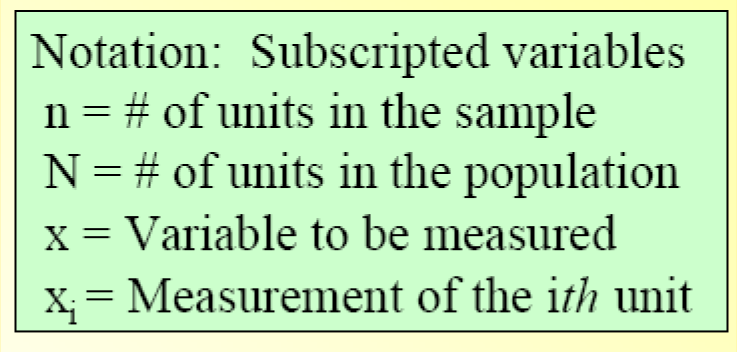


Mean- Arithmetic Average

Mean of a Sample - $\bar{x}$

Mean of a Population - **μ**

Can be calculated only on quantitative data



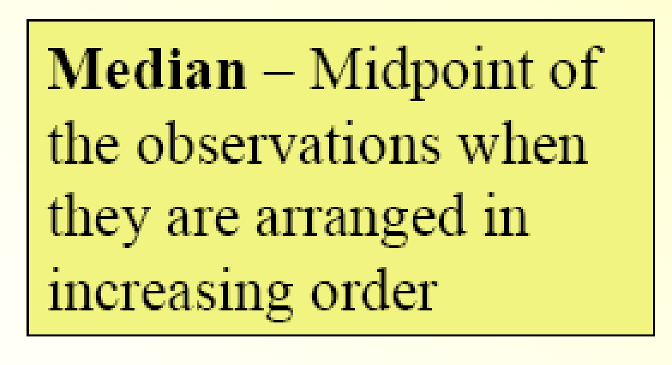
Notation: Subscripted variables

n = # of units in the sample

N = # of units in the population

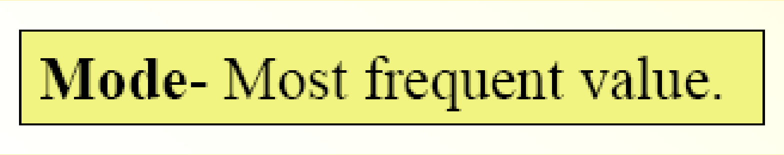
x = Variable to be measured

x_i = Measurement of the *ith* unit



Median – Midpoint of the observations when they are arranged in increasing order

Can be calculated on *quantitative* **or** *ordinal* data



Mode- Most frequent value.

Can be calculated on quantitative, ordinal, or nominal data!

Attendance Survey Question #7

- On an index card
  - Please write down your name and section number
  - Today's Questions: